WHAT IS CLAIMED IS:

1. A method of determining similarity between words, comprising:

receiving as an input a first word and a first dependency structure that includes the first word;

receiving a data structure indicative of a second word and a second dependency structure that includes the second word; and

calculating the similarity between the first and second words based on a similarity measure weighted based on a frequency indicator indicative of a frequency of occurrence of the second dependency structure in training data.

2. The method of claim 1 wherein calculating the similarity further comprises:

weighting the similarity measure using a weighting measure selected from a plurality of different weighting measures based on the frequency indicator.

3. The method of claim 2 wherein the plurality of weighting measures includes a co-occurrence weighting measure and a mutual information (MI) weighting measure.

4. The method of claim 3 wherein weighting the similarity measure comprises:

weighting the similarity measure with the co-occurrence frequency measure if the frequency indicator indicates the frequency of occurrence is below a frequency threshold.

5. The method of claim 4 wherein weighting the similarity measure comprises:

weighting the similarity measure with the MI measure if the frequency indicator indicates the frequency of occurrence is above the frequency threshold.

6. The method of claim 2 wherein receiving a data structure indicative of a second word comprises:

accessing a data store that stores records that include words and associated dependency structures and frequency indicators.

7. The method of claim 6 wherein the associated dependency structures and frequency indicators in the data store are stored as vectors associated with the words, and wherein accessing a data store comprises:

accessing the words and associated vectors.

8. The method of claim 6 wherein accessing the data store comprises:

identifying candidate words in the data store by reducing the search space of records in the data store.

9. The method of claim 8 wherein identifying candidate words comprises:

accessing a lexical knowledge base to identify possible candidate words in the data store.

10. A method of generating annotated data for use in calculating similarity between words, comprising:

receiving a textual input;

parsing the textual input into dependency structures including words and relation types indicative of relations between the words in the textual input;

generating a vector corresponding to each dependency structure, the vector including a related word, a relation type indicator, and a frequency indicator indicating a frequency with which the dependency structure occurred in the textual input; and

storing the words and corresponding vectors regardless of the frequency with which the dependency structures occurred in the textual input.

11. The method of claim 10 wherein the frequency indicator comprises a normalized count value.

12. The method of claim 10 wherein parsing the textual input into dependency structures comprises:

parsing the textual input into dependency triples.

13. A natural language processing system, comprising:

a data store storing head words and associated attributes, each of the attributes including a related word that was related to the head word in a training corpus, a relation type indicator indicating a type of relation between the head word and the related word, and a frequency indicator indicative of a frequency with which the attribute occurred relative to the head word in the training corpus; and

a similarity generator configured to receive an input word and an associated input dependency structure and to access the data store and calculate a similarity between the input word and head words in the data structure based on the input word and associated input dependency structure and the head words and associated dependency structures using a similarity measure that weights a similarity corresponding to a given head word based on the frequency indicator associated with the given word.

14. The system of claim 13 wherein the similarity generator is configured to weight the similarity with a weighting measure.

15. The system of claim 14 wherein the similarity generator is configured to select one of a plurality of weighting measures to weight the similarity based on the frequency indicator associated with the given head word.

16. The system of claim 15 wherein the similarity generator is configured to select a co-occurrence frequency weighting measure if the frequency indicator is below a predetermined threshold.

17. The system of claim 16 wherein the similarity generator is configured to select a mutual information weighting measure if the frequency indicator is above the predetermined threshold.

18. The system of claim 13 and further comprising:

a lexical knowledge base, the similarity generator being configured to access the lexical knowledge base to identify a subset of the head words in the data store as candidate words prior to calculating the similarity.

19. The system of claim 13 wherein the data store stores the attributes as vectors.

20. A system for generating annotated data for use in calculating similarity between words, comprising:

a parser configured to receive a textual input and parse the textual input into dependency structures including words and relation types indicative of relations between the words in the textual input and generate a vector corresponding to each dependency structure, the vector including a related word, a relation type indicator, and a frequency indicator indicating a frequency with which the dependency structure occurred in the textual input; and

a data store configured to store the words and corresponding vectors regardless of the frequency with which the dependency structures occurred in the textual input.

21. The system of claim 20 wherein the frequency indicator comprises a normalized count value.

22. The system of claim 20 wherein the parser is configured to parse the textual input into dependency triples.